**Exploration about the data**

Variable Name Description

Survived Survived (1) or died (0)

Pclass Passenger’s class

Name Passenger’s name

Sex Passenger’s sex

Age Passenger’s age

SibSp Number of siblings/spouses aboard

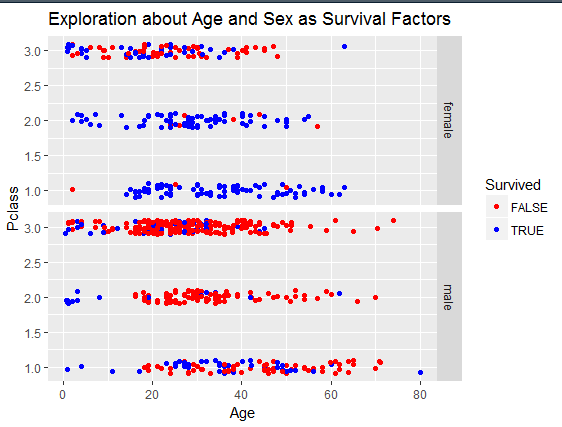
Parch Number of parents/children aboard

Ticket Ticket number

Fare Fare

Cabin Cabin

Embarked Port of embarkation



Through this image, we can find that age and sex are very important factors.

**Data preprocessing**

1. create new feature(colum) family size

It is better to combine the feature” SibSp” and “Parch”, because they describe the situation of families

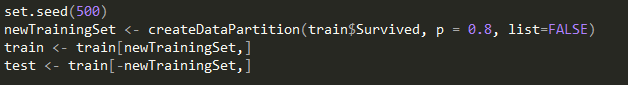
C:\Users\HMACCE~1\AppData\Local\Temp\1519010332(1).png

1. sensible value imputation

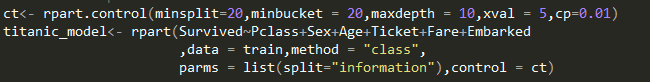
Through the csv file, we can find that there are null value in the column “Age”. Because “Age” is very important factor, we must fill in missing age value

C:\Users\HMACCE~1\AppData\Local\Temp\1519010351(1).png

1. Divide the data into the training set and testing set



**Decision Tree models**



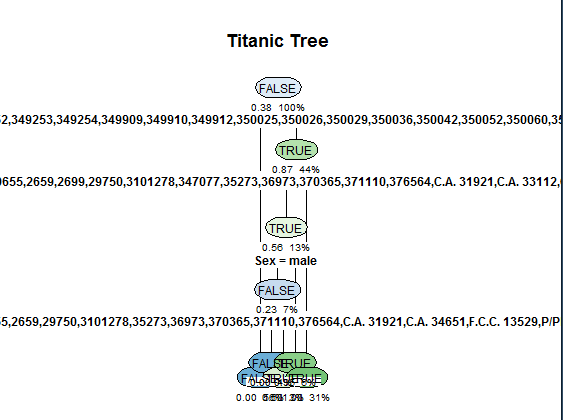
Params Description

Survived~Pclass+… formula

Method=”class” if y is a factor then method = "class" is assumed

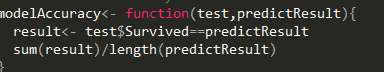
Spilit=information The splitting method is the gain of the information

Besides, I drawled a picture of the decision tree



**Evaluation of models**

In order to calculate the accuracy, I write a function:



And the result is the 0.522409.